United States Department of Agriculture

National Agricultural Statistics Service

Research and Applications Division

SRB Research Report Number SRB-89-11

October 1989

MARKOV CHAIN FORECASTS OF COTTON OBJECTIVE YIELD

James H. Matis
Charles R. Perry
Donald E. Boudreaux
Dave J. Aune

MARKOV CHAIN FORECASTS OF COTTON OBJECTIVE YIELD

by James H. Matis*, Charles R. Perry, Donald E. Boudreaux*, and Dave J. Aune, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, October 1989, Research Report No. SRB 89-11.

Abstract

This paper reports an evaluation of a Markov chain procedure for forecasting final cotton objective yield from pre–harvest objective yield measurements. The evaluation was based on 1980 through 1986 data for six major cotton producing states: Arizona, Arkansas, California, Louisiana, Mississippi and Texas. Comparisons between the forecast errors for the Markov procedure and the forecast errors for the current National Agricultural Statistics Service (NASS) procedure showed that, at the six state level, the Markov procedure performed better than the current NASS procedure in August, about the same in September, and worse in October. Since the Markov and NASS forecast errors were based on different data sets, these comparisons should be viewed with some caution. Separately comparing the two sets of forecast errors showed that the mean forecast error of the NASS forecast procedure improved from month to month as the percentage of the objective yield plots harvested increased, whereas the Markov procedure does not show the same month to month improvement. The Markov procedure has advantage over the current procedure in terms of data collection because it does not require the collection of plot–level survival data (tag data). Thus, if the Markov procedure can be modified to have forecast errors similar to the current procedure, its use would greatly reduce the time spent in the field collecting data and thus reduce data collection costs. Suggestions are made for modifying the Markov forecast procedure to ensure that it converges month by month to the final objective yield estimates.

* Dr. Matis and Mr. Boudreaux are with Texas A&M University, Department of Statistics, College Station, Texas 77843.

Keywords: Markov Chain, Cross Validation, Forecast Errors, Objective Yield.

Acknowledgements

The authors express their appreciation to Ben Klugh and George Hanuschak for their helpful suggestions and moral support during this project. We express our gratitude to Jeff Geuder for his assistance in assembling the maturity category data of Appendix E. Finally, we thank our colleagues Bill Donaldson, Barry Ford, and Phil Kott for their thoughtful reviews of this report. However, we bear full responsibility for any errors.

Contents

Summary

This is the third in a series of three joint studies by the National Agricultural Statistics Service (NASS) and Texas A&M University (TAMU) that have applied the statistical theory of Markov chains to forecast crop yields from pre–harvest data. The first study applied the Markov chain approach to simulated corn yield data (Matis *et al.* 1985). The second study applied the Markov chain approach to NASS's corn objective yield survey data (Matis *et al.* 1989).

The current study applies the Markov chain approach to forecast final cotton yield from pre–harvest objective yield measurements. Analyses were conducted using 1980 through 1986 data for each of six major cotton producing states: Arizona, Arkansas, California, Louisiana, Mississippi, and Texas. Cross validation techniques were used to estimate the Markov chain forecast errors. These estimated forecast errors were compared to historical NASS forecast errors. The comparison showed that the Markov chain approach performed better than the current NASS procedure in August, about the same in September, and worse in October. However, these comparisons should be viewed with some caution since the Markov and NASS forecast errors were based on different data sets. One limitation of the Markov approach seems to be that its forecast errors do not tend to decrease from month to month — the Markov procedure did not converge as the growing season progressed to the final at–harvest yield estimate.

However, the Markov procedure has a potential advantage over the current procedure in terms of data collection because the Markov procedure does not require the collection of plot–level survival data (tag data). Thus, if the Markov procedure can be modified to have forecast errors similar to the current procedure, its use would greatly reduce the time spent in the field collecting data and thus reduce data collection costs.

Recommendations for further research are proposed. Some recommendations aim at modifying the Markov forecast procedure to ensure that it converges to the final at–harvest yield estimate. Others are aimed at applying resampling techniques simultaneously to the current NASS procedure (including its two components) and the Markov procedure to ensure that future evaluations are based on comparable forecast errors estimates. Another group of recommendations address the survival modeling of the tag data. It is suggested that Markov and semi–Markov process models be formulated and fitted to the data with the aim of assessing the potential contribution of such data for yield forecasting. The remaining recommendations are aimed at improving the variable selection process, adapting the SAS software for PC processing, adapting the Markov procedure software to provided current to historical years similarity measurements, and applying the methods to other crops.

Introduction

The National Agricultural Statistics Service (NASS) and Texas A&M University (TAMU), Department of Statistics, have applied a Markov chain theory to forecast crop yield from early season crop data in two earlier studies. Markov chain theory is the study of time developing processes which assumes that all the current information about the future development of a process (final objective yield) is contained in knowledge of the current

state of the process (current objective yield measurements). The general procedure has been outlined in previous papers in the literature. The first (Matis, Saito, Grant, Iwig and Richie, 1985) outlined the basic Markov chain approach and applied it successfully to simulated corn yield data. The second (Matis, Birkett and Boudreaux, 1989) adapted the procedures to the large–scale USDA corn objective yield survey. Another (Grant, Matis and Miller, 1988) applied the approach to forecasting the commercial shrimp harvest in the Gulf of Mexico.

This technical report is a continuation of the research in this area, and has the following four specific objectives:

1. To investigate possible improvements in the basic Markov chain approach,
2. To develop a computer algorithm to calculate and set the Markov forecasts and thus eliminate any human bias,
3. To implement the revised procedures and new computer program for all six major cotton producing states using all available survey data (1980–86), and
4. To compare the results from the Markov chain approach to current operational methods.

The results are presented in separate subsequent sections with future research recommendations.

Improvements in the Basic Markov Chain Approach

This report evaluates a change in the way crop condition categories are defined. These categories constitute the states of the Markov chains, and hence their definition and construction is an integral step in the approach. In the previous articles, the states were constructed using a two–way factorial arrangement. A primary predictor variable, say X_1, and a secondary predictor variable, say X_2, were subdivided separately into n_1 and n_2 classes, respectively. The two sets of classes were then crossed to give $n_1 \times n_2$ combined condition classes.

The previous procedure is relatively simple to implement. The two predictor variables are chosen on the basis of their maximum predictability. In practice the two variables are always correlated, sometimes highly correlated. As a consequence the factorial construction procedure, which utilizes only the marginal distributions of the two variables, leads to many categories which are virtually empty. Such sparse categories in turn reduce the efficiency of the forecasting procedures.

The new, modified procedure takes into account the possible correlation of the two predictor variables. The new procedure is to first subdivide the primary predictor variable into n_1 classes. Then to subdivide the secondary predictor variable into n_2 classes within each of the n_1 classes of the primary variable. This new procedure would in principle equalize the number of observations of the baseline data in each of the $n_1 \times n_2$ cells. This obviously requires a more difficult computer algorithm. The algorithm is further complicated by the fact that the secondary variable is often discrete and is not readily divisible into n_2 categories of equal size.

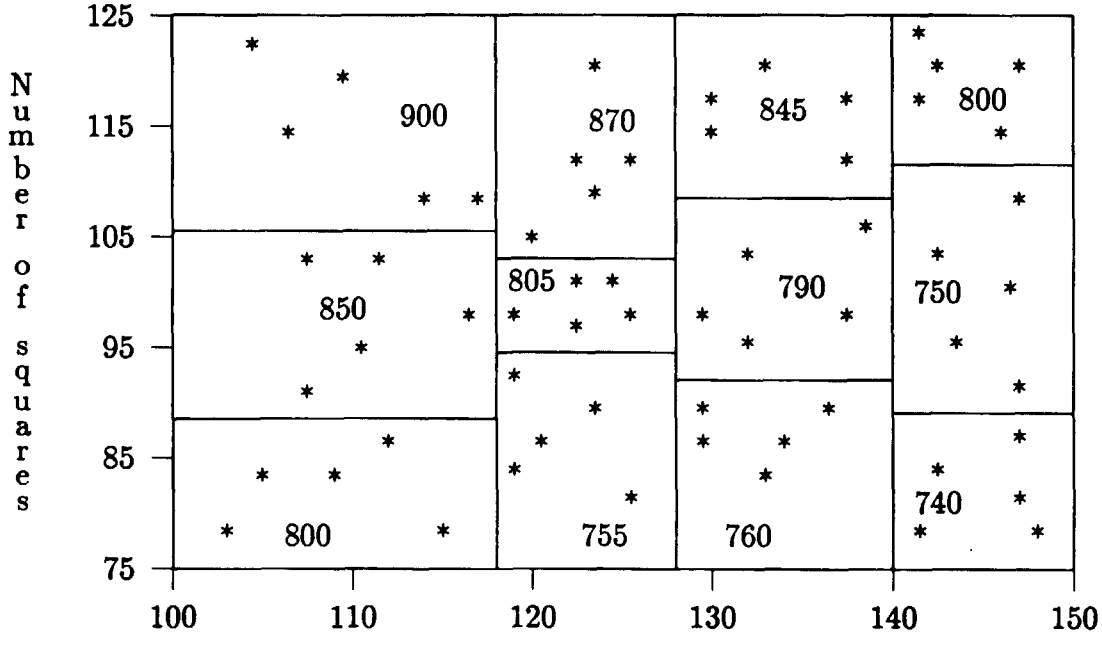
The example that follows, which is an oversimplification of the actual Markov procedure, demonstrates the basic ideas. Several intermediate steps which add richness and complexity have been omitted. However, the general idea underlying the final application of the new, modified Markov procedure is preserved.

Let us say that we have historic data from 60 cotton objective yield plots. The Markov forecast model is derived in four steps.

1. Determine the best two predictor variables. For the example we will use number of plants as the best predictor, X_1, and number of squares as the second best predictor, X_2.
2. Divide the range of the best predictor into four intervals such that each interval has the same number of observations. We have assumed that n_1 equal four.
3. Divide the range of the second best predictor into three intervals such that each interval has the same number of observations. We have assumed that n_1 equals three. This process has divided the 60 plots into 12 cells.
4. Calculate the average historic final yield of the plots in each cell.

Illustration 1 summarizes the example. The numbers inside the cells are the average final yields as pounds per acre. To use the table for current year data, take the number of squares and the number of plants from a plot and find cell in which it falls. The forecasted yield for the plot is the average historic final yield of the given cell. For example, if a plot from this year's survey has 100 squares and 125 plants, then the forecasted yield for the plot is 805 pounds per acre.

Illustration 1. Forecast Plot Yield by Cell.



The principal conceptual difference between the example and the actual Markov chain procedure is that in the actual procedure the states are defined for every month for which data are gathered. Transition probability matrices are estimated for each month of the growing season. Each monthly matrix gives for each state of that month the probabilities associated with each state in the subsequent month. The monthly transition probability matrices are combined to obtain the Markov chain probability model which yields a weighted average historical final yield, as illustrated in the example.

Revised Computer Programs

The existing computer programs were revised and expanded to accomplish two objectives. The first is to implement the new procedure for defining states. The second is to automate the program so that the primary and secondary variable are determined for each specific forecast. Previously these predictor variables were selected on the basis of expert judgment of the user. The revised procedures select the variables based solely on the empirical evidence available at the time of each forecast. The revised program is divided into the following five steps:

1. Read and edit the data,
2. Select the independent, predictor variables,
3. Create the new categorical states based upon the selected variables and the user–specified number of subdivisions,
4. Calculate the Markov chain transition matrices, and
5. Generate the forecasts, determine their accuracy and estimate the forecast error.

A brief description of computer programs used to implement the revised procedures is given in Appendix A. For those interested in further details of the computer algorithm, Matis, Perry and Boudreaux (1989) give the complete Statistical Analysis System (SAS) code along with a detailed annotation.

Test of Markov Forecasting Procedures on Cotton Yields

The modified procedures and computer software were tested on all available data (1980–86) for all six major cotton producing states. Cross validation methods (Efron, 1982, Chapter 7) were used to estimate the forecast errors, as in previous studies. Further details of these computations are given in Appendix B. The forecast error was estimated within each cotton–producing state for each year for which data were available. Appendix B contains a detailed description of data availability by state. For each year in question, the model for a particular state was fitted to data for the remaining years to see how well the fitted model predicts the excluded year. The details of model development are given in Appendix C, which contains a listing of all variables used in the study and a table of the primary and secondary predictor variables by state, year, and month. The predictor variables chosen by the computer algorithms are fairly consistent from year to year within the states but vary substantially between the states. For example, the first two states, Arkansas and California, have one variable in common (total boll count) which is selected

in August each year. However, the other variable selected, number of squares in Arkansas and number of large bolls in California, are different between the two states but are always the same within the states. Referring to the maturity category tables given in Appendix E, one observes that cotton in Arkansas is always in an earlier stage of development than cotton in California; thus, the variables selected tend to be reasonable. Table 1 contains the resulting percent forecast error by state, year, and month. Operationally, California and Texas are subdivided into two regions; however the data were too sparse to subdivide California in the analyses. No data were available for the 1982 harvest in Mississippi. The mean (absolute) percent errors from all available data in the six states were 2.9% for Arizona, 17.6% for Arkansas, 4.5% for California, 23.61% for Louisiana, 14.6% for Mississippi, 13.7% for Texas, Region 1, and 17.2% for Texas, Region 2.

Table 1

% Forecast Error by State, Year and Month.
(Actual Yields Given in Parentheses). 1980–1986 Data.

State 4 — Arizona

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(916.5)	(964.0)	(942.1)	(873.3)	(923.6)	(977.6)	(959.0)	
August	4.2	2.6	-4.5	-5.7	-0.2	-2.8	1.0	3.0
September	4.8	3.8	-2.9	-3.8	1.1	-2.0	2.2	2.9
October	6.2	3.8	-0.9	-1.8	3.1	-1.1	2.6	2.8
Mean	5.1	3.4	2.8	3.8	1.5	2.0	1.9	2.9

State 5 — Arkansas

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(205.5)	(323.9)	(368.9)	(289.8)	(402.6)	(424.5)	(372.7)	
August	63.0	-0.2	3.2	-1.9	-27.2	-19.9	-2.4	16.8
September	64.1	1.3	3.9	0.1	-26.7	-18.2	-1.7	16.6
October	66.4	3.7	4.7	5.9	-24.7	-18.2	-1.7	17.9
Mean	64.5	1.7	3.9	2.6	26.2	18.8	1.9	17.6

Table 1 (Continued)

State 6 — California

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(747.3)	(894.9)	(867.5)	(791.6)	(775.6)	(922.2)	(875.8)	
August	5.0	-0.4	-8.8	-7.3	9.5	-1.5	0.5	4.7
September	5.4	0.3	-7.2	-6.3	-10.5	-1.0	0.9	4.5
October	6.5	0.4	-5.5	-4.9	11.2	-0.9	1.1	4.4
Mean	5.6	0.4	7.2	6.2	10.4	1.1	0.8	4.5

State 22 — Louisiana

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(255.4)	(352.6)	(476.5)	(208.9)	(540.8)	335.3)	(409.8)	
August	58.2	12.9	-19.5	12.2	-42.5	18.2	0.0	23.4
September	59.9	14.7	-19.0	13.4	-41.3	18.8	0.0	23.9
October	61.5	15.1	-13.4	15.8	-39.5	19.8	0.6	23.7
Mean	59.9	14.2	17.3	13.8	41.1	18.9	0.2	23.6

State 28 — Mississippi

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(338.0)	(455.8)		(350.8)	(169.0)	(440.3)	(374.7)	
August	15.9	-6.0	no	10.0	-33.6	-0.5	19.2	14.2
September	16.7	-5.6	data	11.5	-33.0	0.3	19.7	14.5
October	17.8	-5.2		14.5	-33.2	1.3	20.3	15.4
Mean	16.8	5.4		12.0	33.3	0.7	19.7	14.6

Table 1 (Continued)

State 48 — Texas, Region 1

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(191.7)	(268.5)	(246.9)	(349.1)	(227.7)	(354.8)	(228.2)	
August	38.1	-6.6	0.1	-17.1	20.0	-26.6	-2.7	15.9
September	42.5	-4.8	2.7	-12.9	15.0	-19.7	-1.8	14.2
October	42.8	-2.9	-8.7	-12.8	-0.0	-3.8	-6.0	11.0
Mean	41.1	4.8	3.8	14.3	11.7	16.7	3.5	13.7

State 48 — Texas, Region 2

Year	1980	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(178.6)	331.9)	(233.2)	(243.0)	276.9)	(320.3)	(219.4)	
August	62.4	-16.6	0.2	2.2	-10.9	-12.2	13.7	16.9
September	63.4	-16.0	1.4	3.0	-10.2	-11.7	14.4	17.2
October	63.9	-15.3	3.7	4.3	-9.0	-11.2	15.4	17.5
Mean	63.2	16.0	1.8	3.2	10.0	11.7	14.5	17.2

The mean errors were higher than expected, when compared with the previous pilot study (Matis, Birkett and Boudreaux, 1989), for four of the states, Arkansas, Louisiana, Mississippi and Texas. The inflation of errors was due to the inclusion of the 1980 data. In 1980 cotton yields were substantially lower than in any other years in these four states. With the lack of such a historical precedent, the mean error percentages for 1980 were 64.5% for Arkansas, 59.9% for Louisiana, 16.8% for Mississippi and 41.1% and 63.2% for the two regions in Texas. In order to study further the effect of the unusual 1980 season, the data in Region 2 in Texas were reanalyzed after deleting the 1980 data. The results are given in Table 2. The overall results improve dramatically, with a new mean error rate of 10.3%, as compared to the previous 17.2%.

Table 2

% Forecast Error by Year and Month for Texas, Region 2
(Actual Yields Given in Parentheses) 1981-1986 Data.

Year	1981	1982	1983	1984	1985	1986	Mean
Actual Yield	(331.9)	(233.2)	(243.0)	276.9)	(320.3)	(219.4)	
August	-12.0	5.7	8.8	-5.9	-7.2	21.1	10.1
September	-11.4	6.19	9.16	-5.2	-6.6	21.9	10.3
October	-10.8	9.4	10.1	-3.9	-6.2	22.7	10.5
Mean	11.4	7.3	9.5	5.0	6.7	21.9	10.3

Figures 1–3 illustrate the reasons for the poor forecasting results in Region 2 of Texas in 1980. Through an analysis of the 1981–86 data, the best predictor variable, X_1, in August of the final yield was found to be the number of squares, which is coded as X3-8. Figure 1A is a scatterplot for the combined 1981-86 data of yield *vs.* the number of squares in August. Figure 1B is the corresponding scatterplot for the 1980 data. The distribution of the number of squares was subdivided into four classes, and the mean yield calculated within each quartile. It is clear that the prediction based on the 1981–86 data would be much higher than the actual 1980 realization regardless of how one might partition the number of squares.

Figures 2 and 3 present the corresponding data on which the September and October forecasts for Region 2 of Texas are based. The second step of the revised computer program analyzed the 1981–86 data to determine which variables in September and October were the best predictors of final yield. The primary variable selected for both months was the number of large bolls, which is denoted as LB-9 and LB-10 for the two periods. The scatterplots of yield *vs.* these variables for 1981–86 are given in Figures 2A and 3A. The scatterplots for the 1980 data of yield *vs.* these independent variables are given in Figures 2B and 3B. Note that the distributions of large bolls in September and October of 1980 are similar to the 1981–86 distributions. However the forecasts derived from the 1981–86 data would substantially exceed the 1980 realizations, as noted previously in Figure 1.

The current Markov chain procedures differ conceptually from the graphical approach demonstrated with Figures 1–3 in two minor ways. One is that the predictor variables are chosen by a nonparametric rank regression procedure, by regressing the ranks of the response variable on the ranks of the predictor variables to determine the best pair of predictors. Another is the concurrent use of a second, correlated predictor variable. Neither of these differences would alter the basic problem with the 1980 data, namely that the plant growth characteristics appeared to follow a normal pattern in August through October, yet the final yields were far below expectations due presumably to unexpected weather (or economic) conditions.

Figure 1. Plot Yield *vs.* Number of Squares in August

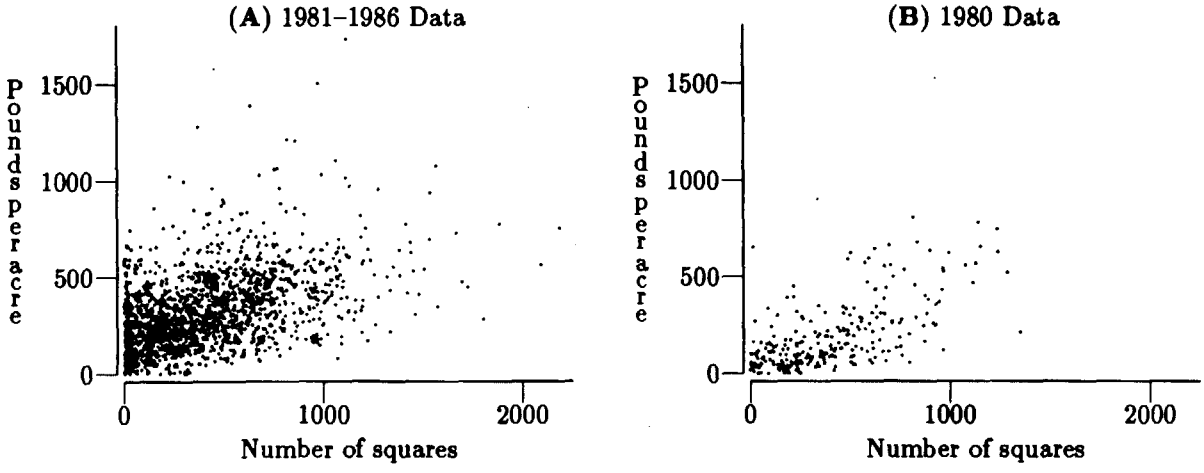


Figure 2. Plot Yield *vs.* Number of Large Bolls in September

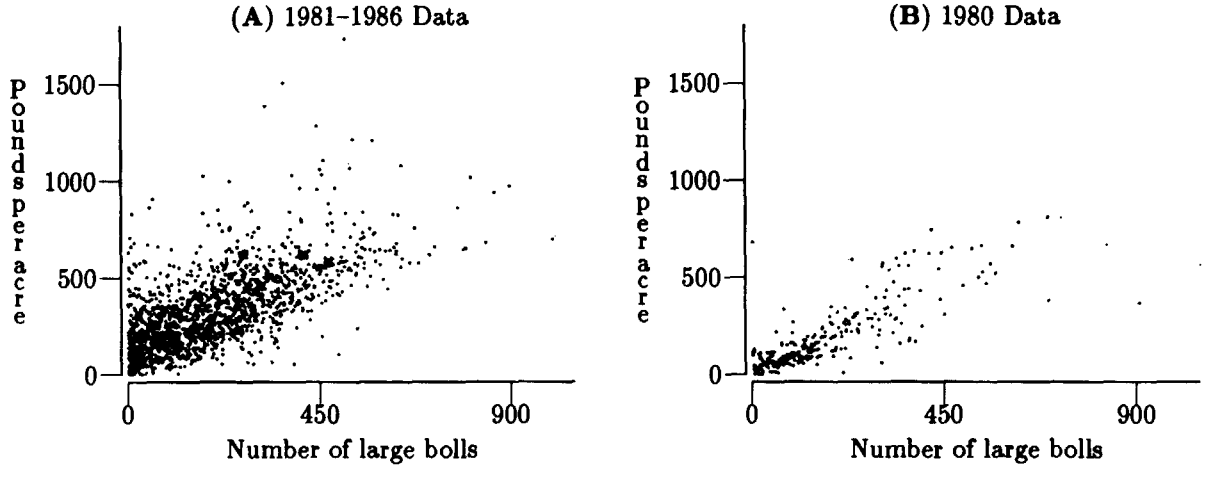
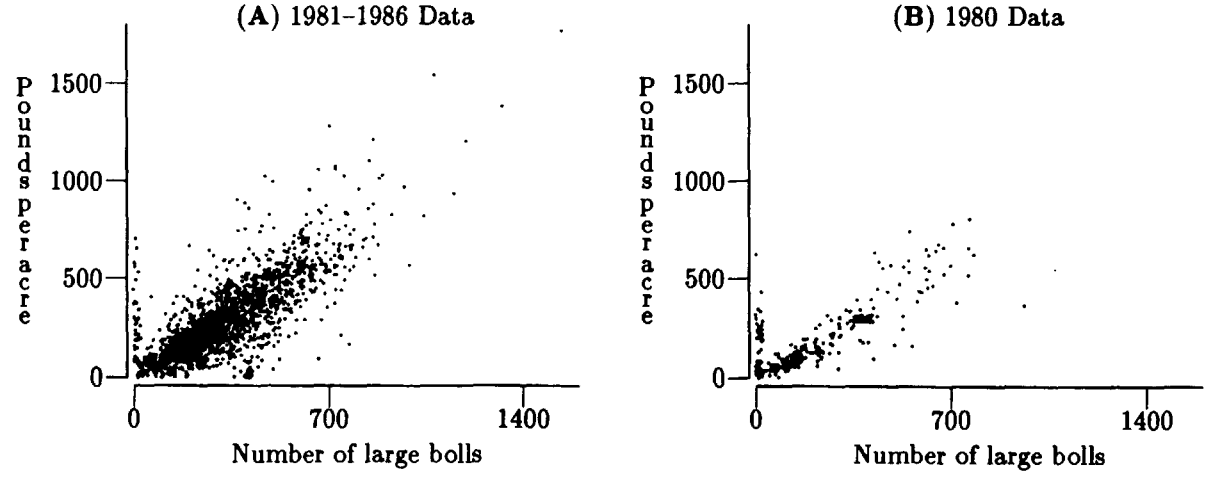


Figure 3. Plot Yield *vs.* Number of Large Bolls in October



It seems apparent that no forecasting methodology which utilizes only the current variables could successfully predict the 1980 yields. One possible solution would be to include as an independent variable some other measurable characteristic that is a leading indicator of future conditions but largely independent of the present variables. One such available variable is the "pasture and range condition" variable which is recorded monthly in each of the states of interest. It would be easy to include the variable in the Markov chain procedures. Another possible solution is to transform the dependent variable, as will be described in a later section.

Comparison of Test Results with Current NASS Procedure

Rigorous comparisons of the forecasts given by current USDA procedures and the forecasts generated from the current Markov chain procedures are not possible since the necessary historical data are not available. However a limited comparison was made by comparing the Markov chain cross validation results to the USDA/NASS operational results for the years 1982–1986. This means that the Markov chain forecasts were based on all data from the years 1980–1986 other than the year in question while the operational forecasts were based on the previous five years. Two sets of somewhat offsetting biases are probably inherent in these comparisons make them hard to interpret. First, the errors associated with the Markov procedure are probably biased downward since the data used in deriving all but the last Markov resample estimate contain both leading and trailing data. Second, the errors associated with the Markov chain forecast includes the effect of 1980 which gives an upward bias. This second type of bias may have been removed from the errors associated with the operational procedure by operator (human) intervention to remove outliers.

A number of observations can be drawn from this research. The conclusions that involve comparisons between the Markov and NASS forecasts should be interpreted in light of the reservations discussed in the above paragraph. However, the conclusions that do not involve comparisons between the Markov and NASS forecasts, *i.e.* the internal comparisons of the Markov (or NASS) procedures over months, are not limited by the discussion in the above paragraph.

1. Conclusions that involve comparisons between the Markov and current NASS procedure results (see Table 3). The validity of these comparisons may be limited by the lack of comparability of the operational and cross validation results.
   a. In August the mean square errors from the Markov procedure are much smaller than the mean square errors from the current NASS procedure at the six state level. However, in two of the six states, the mean forecast error of the Markov procedure are slightly larger.
   b. In September the mean square errors from the Markov and the current NASS procedures tend to be similar at the six state level. For Texas, the Markov procedure performs substantially better than the operational procedure. However, for Mississippi, the Markov procedure performs much worse than the operational procedure.

c. In October the mean square errors from the Markov procedure are larger than the mean square errors from the current NASS procedure at the six state level. For Louisiana and Mississippi, the mean forecast error of the Markov procedure are substantially larger than those of the operational procedure.

d. It should be observed that the Markov model does not require plot–level survival measurements (tag data) nor a maturity category determination.

Table 3.

Comparison of Mean % Forecast Error for 1982–1986
by State and Month

	STATE	FORECAST	AUG	SEPT	OCT
1.	Arizona	NASS	5.6	2.8	3.8
		Markov	2.8	2.4	1.9
2.	Arkansas	NASS	13.8	13.3	13.8
		Markov	10.9	10.1	11.0
3.	California	NASS	3.8	5.1	1.5
		Markov	5.5	5.1	4.7
4.	Louisiana	NASS	17.7	13.7	4.4
		Markov	18.5	13.1	17.8
5.	Mississippi	NASS	17.8	8.3	4.9
		Markov	15.8	16.1	17.3
6.	Texas	NASS	12.7	8.8	4.2
		Markov	2.9	2.0	2.8
	Weighted Mean	NASS	11.3	8.2	4.5
		Markov	7.1	6.1	6.9

2. Conclusions that involve only internal comparisons of the Markov or current NASS procedure results.

a. The mean square error of the current NASS procedure tends to decrease from month to month. For example, the weighted average over the six states of the mean square errors of current NASS forecasts were 11.3, 8.2, and 4.5, respectively, for August, September and October (see Table 3). The convergence of the cur-

rent NASS forecasts to the final objective yield is probably explained by the fact that as the season progress the form of the equations used in the current NASS procedure (and the data going into them) tend to converge to the final objective yield equation (see Section 8, Forecasting and Estimating Models, of the Objective Yield Supervising and Editing Manual which is included in this report as Appendix D).

b. The mean square error of the Markov procedure does not tend to decrease from month to month. For example, the weighted average over the six states of the mean square errors of current NASS forecasts were 7.1, 6.1 and 6.9, respectively, for August, September and October (see Table 3). The lack of convergence of the Markov forecasts to the final objective yield forecasts is probably due to the fact that, unlike the equations of the current NASS procedure, the Markov forecast does not make use of the current year data other than through counts for the two predictor variables. In other words, at the cell level the Markov forecasts are just the mean historical yield associated with the level of the variables defining the cell.

Conclusions

In this report forecast errors were computed for the Markov chain procedure using cross validation methods by state and region for 1980 to 1986. These forecast errors for gross yield were then compared to historical NASS forecast errors. Thus the forecast errors associated with the Markov forecast and the current NASS forecast procedure are based on different historical data. Hence the comparisons made between the two procedures do not necessarily provide reliable estimates of expected operational results. (The original plan was to compute forecasts and errors for both the Markov and current NASS procedures by cross validation techniques. However, cuts in anticipated research funding precluded the programming necessary to compute cross validation forecast errors for the current NASS forecast procedure.) The principal findings of this report follow:

1. The new method of selecting predictor variables was implemented. The set of predictor variables chosen by the Markov procedure varied from state to state and from one part of the growing season to another. However, for a given forecast month the predictors chosen showed little year to year variation within states and regions. In most cases the prediction were associated with the number of fruiting bodies or the weight of seed cotton harvested – however six percent of the time the second predictor chosen was the number of plants.
2. A set of computer algorithms which eliminate all operator intervention was produced and successfully implemented.
3. In August, the mean forecast error (MFE) from the Markov procedure was smaller than the MFE from the current NASS procedure; in September, they were roughly equal; but, in October, the MFE from the Markov procedure was larger than the MFE from the current NASS procedures. These comparisons should be viewed with some caution since the Markov and NASS forecast errors were based on different data sets.

4. The Markov algorithm does not depend on any plot–level survival data (tag data). This means that the data collection effort to exercise the Markov procedure would be much less costly and time consuming than that for the current procedure.

Recommendations for Further Research

Markov chain procedures are very flexible and have been successfully adapted previously for forecasting crop yields and commercial fisheries harvests. These procedures can be modified in a number of ways to overcome the structural problems discussed in the earlier sections of this report. In all, 14 recommendations for further research are listed below.

The first eight recommendations, which are listed under Part 1, are concerned with modifying the basic Markov chain procedure and comparing the modified procedure with the current NASS procedure. Highest priority should be given to these seven recommendations in future research. Three other closely related areas of general research are recommended, Parts 2–4. The recommendations listed under Part 2 address the survival modeling of the tag data. They suggested that Markov and semi–Markov process models be formulated and fitted to the data with the aim of assessing the potential contribution of such data for yield forecasting. The recommendations listed under Parts 3–4 are aimed at improving the variable selection process, adapting software, and applying the methods to other crops. The research listed under Parts 2–4 should be evaluated separately from Part 1. One or more of the last three parts can be investigated concurrently with Part 1.

Part 1.

The main recommended modification, which relates to the dependent variable being predicted, is aimed at modifying the Markov forecast procedure to ensure that it converges to the final at–harvest yield estimate. It is described first along with a rationale for its likely success. Other recommendations in this group of future research recommendations are aimed at applying resampling techniques simultaneously to the current NASS procedure (including its two components) and the Markov procedure to ensure that future evaluations are based on comparable forecast errors estimates. Still others are aimed adapting the SAS software for PC processing, adapting the Markov procedure software to provided current to historical years similarity measurements.

1. Even though the Markov procedure appears to be some what superior to the current NASS procedure early in the season, it still has two major weaknesses. The Markov forecast, unlike the current NASS forecast, neither tends to improve as the season progresses nor tends to converge to the final objective yield estimate at the end of the growing season.

   Combining the current NASS forecast equations for boll number and average boll weight and evaluating the parameters in the resulting equation at their idealized at–harvest values produces essentially the final objective yield estimation equation. In other words, as the growing season progresses the combined NASS forecast gets closer

to the final objective yield estimate. In contrast, since the Markov procedure always relies on historical yield values in making its forecast, the equation associated with its final forecast does not necessarily approach the final objective yield estimation equation. In other words, the final Markov forecast is always the average of historical plot yield values and thus may never approach the average of current plot yield values.

To overcome the difficulties referred to above, the following modification to the current Markov procedure is recommended. The basic idea is to use the Markov procedure to forecast the gross yield remaining in the plot after the current data collection period; and then combine this forecast with the cumulative enumerator harvested yield to obtain an at-harvest gross yield forecast. This modified Markov procedure will converge to the final objective yield estimate as one moves through the growing season and as a larger percent of the plot is harvested by the enumerator. It appears that the forecast error of such modified Markov procedure would be smaller than the forecast error of either the current NASS forecast procedure or the investigated Markov procedure. This procedure is outlined in detail in Appendix F.

Two other possible modifications are:

2. The unexpectedly low 1980 harvest might have been successfully predicted by using other leading indicators, such as USDA pasture and range condition data. Pasture conditions were very poor in 1980, and such additional variables are easy to incorporate into the Markov chain methodology.

3. The problem with late season forecasting might be improved by a separate analysis with fewer intermediate steps. This modification is simple and at worst would only increase the computer runtime.

One hindrance in the present studies was the mainframe computer implementation of the procedures. It is therefore recommended that consideration be given to the following items.

4. The program should be modified for use on a microcomputer. The current SAS code would need some adaptation for efficient production usage for use on USDA and TAMU microcomputers.

Besides improving the Markov chain procedure, it is suggested that the improved procedure be compared to current NASS procedures as follows:

5. An evaluation of Markov forecast procedures (and for that matter any other forecast procedure) should be made using resampling comparisons to an automated version of the current NASS procedure. That is, a computer program should be used to duplicate the current NASS interactive procedure, draw a bootstrap sample, compute the estimate for the current NASS procedure and the investigated procedure without human intervention, and then compare the forecast errors. Since this type of evaluation bases both sets of estimates on the same set of data, it can be expected to provide reliable estimates of the relative magnitude of the forecast errors associated with the two procedures even when the resampled estimates of forecast errors are biased.

6. The evaluation should include not only an assessment of the forecast errors relative the current NASS forecast procedure, but also an assessment of the forecast errors relative to its two component parts. (The current NASS forecast is a composite of a regression model forecast and a survival model forecast.)

7. The set of comparisons among the Markov, NASS automated, and NASS operational procedures that are based on the special bootstrap replicates corresponding to taking six consecutive years of data, using the first five for model development and the sixth for error estimation should be presented separately. Comparisons between the forecast errors, the regression coefficients, etc. from the automated and the operational NASS procedures over this special set of replicates should be of value in accessing how nearly the automated procedure "mimics" the operational procedure. However, the conclusions based on this special set of replicates will likely be somewhat limited since it will contain only five replicates, assuming 1980–1989 objective yield data is available for the analyses.

8. The software for the Markov procedure should be modified so that it produces a measure of the similarity between the current year and the various historical years. A measure of how dependent the current Markov forecast is on each of the historical years used in its derivation should be of assistance to the National Agricultural Statistics Board in making the expert judgments inherent in setting official production forecasts.

Part 2.

This group of future research recommendations concerns the development of Markov process models to describe the tag survival data and to incorporate it into the Markov chain forecast. Plant researchers are now using differential equations models to describe the growth and longevity of various plant parts. A previous study (Saito, 1985) developed Markov process models for the growth and development of the cotton plant. The study demonstrated the feasibility, based on the plant process simulation data in Matis *et al.* (1985), of using such models to predict final cotton yield. Since that project, Matis and coworkers have developed new semi–Markov models (see *e.g.* Seber and Wild, 1989, Chapter 8) and new analysis procedures and software (Allen and Matis, 1989) which could be used to fit the new models to plot–level survival data (tag data). These considerations lead to the following three recommendations:

9. New Markov process and semi–Markov models should be developed and fitted to the tag survival data.

10. These models should be compared to the Markov chain forecast model, which is not based on tag data, and to the current NASS procedure. Such comparisons would enable researchers to evaluate the potential contribution of the tag data to optimal harvest prediction.

11. If semi–Markov models successfully fit the tag data, then such survival models should be combined with the other forecast procedures.

Part 3.

This group of future research recommendations addresses major extensions of the present software and new generalizations of the Markov chain procedures. Two specific recommendations along these lines are:

12. The software should be modified to select the predictor variables internally. At present, this is a separate step following carefully defined decision rules.
13. Multivariate procedures could be investigated for use in generating efficient predictor variables.

Part 4.

This recommendation suggests an investigation of the use of these forecast procedures for other crops. The final recommendation is:

14. The Markov chain procedures should also be applied experimentally to other crops where forecasting has been challenging.

References

Allen, D.M. and Matis, J.H. (1989). *Practical Biological Modeling and Introduction to KMOD.* Short course sponsored by Biometric Society. March 12, 1989. Lexington, KY.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Grant, E. W., Matis, J. H. and Miller, W. (1988). Forecasting commercial harvest of marine shrimp using a Markov chain model. *Ecological Modelling* 43:183–193.

Matis, J. H., Saito, T., Grant, W. E. Iwig, W. C., and Richie, J. T. (1985). A Markov chain approach to crop yield forecasting. *Agricultural Systems* 18:171–187.

Matis, J. H., Birkett, T., and Boudreaux, D. (1989). An application of the Markov chain approach to forecasting cotton yield from surveys. *Agricultural Systems* 29:357–370.

Matis, J.H., Perry, C.R., and Boudreaux, D.E. (1989). *A Computer Algorithm for Markov Chain Forecasts of Cotton Objective Yield*, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, Research Report No. SRB 89–12.

Saito, T. (1985). *An Application of Stochastic Compartmental Theory to Modeling Plant Growth and Crop Yield.* Ph.D. Dissertation. Texas A&M University, College Station, TX.

Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression.* Wiley, New York, NY.

Appendix A – Program Overview

The process of constructing Markov forecasts was accomplished in a five step process:

1. Read and edit the data.
2. Select independent variables.
3. Create new categorical variables based upon the selected variables and user defined number of breaks.
4. Calculate the Markov transition matrices.
5. Generate the estimated forecasts and check the accuracy of the simulation.

These steps were executed within three Statistical Analysis System (SAS) programs: XDAT (step 1.), XRSQ (step 2.), and YRXX (steps 3., 4., 5.). These programs are briefly described below in terms of their basic purpose, an overview of the procedure and a few comments. The complete SAS code for these programs along with a detailed annotation is given by Matis, Perry and Boudreaux (1989).

It should be pointed out that these programs were written for researching the feasibility of utilizing the Markov procedure for USDA data. Thus the programs are not necessarily optimally coded nor are they intended for "production" use. Also, it should also be pointed out that these programs would have to be updated for use with PC–SAS or newer versions of SAS running on a mainframe computer. This fact is especially evident with the replacement of PROC MATRIX with PROC IML.

Program : XDAT

Purpose:

This program edits USDA produced SAS data sets (from tape), create new variables used in later analysis, and develop a data structure appropriate for Markov analysis.

Procedure:

The original tape contained a series of SAS data sets each representing a single year. These data sets were read and a set of procedures were run on them to help identify an individual plots sequence of information (create a "month" variable) and a unique id number. Then the data was combined, edited, and new variables were created. After that, a data set was created and a merge performed in order to create an overall cumulative yield variable. Certain independent variables were then ranked and a set of data steps were executed in order to organize (and rename) the independent variables by month (8,9, and 10). Finally, these monthly data sets were recombined and a SAS data set was created on a mainframe disk pack.

Comments:

The creation of the month and id variables was necessary because of the way the data needed to be cumulated and structured.

Program : XRSQ

Purpose:

This program uses the data created in XDAT to select the two best variables for predicting yield.

Procedure:

Given the nature of the study, one specific year will always be excluded to allow the remaining data to be used to simulate a prediction. Then the SAS procedure PROC RSQUARE is run on the rank data for each of the monthly time frames (8,9, and 10). The resulting set of "best" two predictor variable is used later in the Markov process as the monthly state variables.

Comments:

This process provides for an objective methodology of variable selection. However, it does assume that an "acceptable" set of independent variables are utilized.

Program : YRXX

Purpose:

This program takes the data from the program XDAT along with the selected variables from the program XRSQ and performs several Markov analysis forecast simulations.

Procedure:

The data is broken into two data sets one with the data to provide the forecasts and the other with the "actual" yield values. Both of these data sets have new variables added to them based upon the selected variables, a set of user defined categories, and the breakpoints of the variable categories for the forecasts data set. These new categorical variables are then formed into Markov transition matrices and used to forecast the "actual" yields.

Comments:

This SAS program was developed to provide for a specific research need – it is not optimally coded and should not be used for production work. Parts of the code are dependent on the version of SAS used; for example, PROC MATRIX will need to be replaced with PROC IML in most cases.

Appendix B – Description of Data and of Cross Validation Study

The initial study by Matis *et al.* (1989) was based on four years of data (1981–84) for two states, Texas and California. The cross validation method determined the average of the prediction errors for each state by leaving out one year at a time and developing a prediction model based on the remaining years. However, the predictor variables were based in part on expert judgment and in a strict sense were not completely determined internally for each state, year and month combination.

The data for the present study included the six major cotton producing states for the 1980–1986 period. Texas was subdivided into two regions, but the data in California were not sufficient, particularly in 1984, to warrant a similar partitioning. No data were available for Mississippi in 1982. Within these constraints, a cross validation procedure was implemented. Unlike the previous study, explicit rules for choosing predictor variables were formulated before any data was analyzed, thereby making the choice of such variables completely independent of any subjective (expert) judgment. Thus, for each year in question, a model was developed exclusively from data from within the given state for the remaining years. The number of such remaining years was either five or six. As before, the prediction from the fitted model was compared to the actual value for the year in question in order to determine the forecast error. A smaller follow–up study investigated the Markov chain procedure for Region 2 of Texas for only the 1981–1986 period. A similar cross validation study was used.

Appendix C – Variables Used for Markov Forecasts

Table C1

Variables Available from the Objective Yield Survey

1.	Current number of squares	TOTSQ
2.	Current number of small bolls and blooms	TOTBM
3.	Current number of large unopened bolls	BOLLUN
4.	Current number of partially opened bolls	BOLLPT
5.	Cumulative number of burrs and bolls (accumulated over all visits to data)	BOLLOP
6.	Cumulative number of burrs and bolls on ground (accumulated over all visits to date)	BOLLGR
7.	Cumulative number of bolls in sample (BOLLUN+BOLLPT+BOLLOP+BOLLGR)	TOTBL
8.	Cumulative weight of harvested bolls (all opened bolls and bolls on the ground are harvested at each visit)	CUMWT
9.	Cumulative average weight per boll	WTBOLL
10.	Number of plants	PLT
11.	Row spacing	ROWSP
12.	Yield per hectare (constant X CUMWT/ROWSP)	Y

Table C2

Primary and Secondary Predictor Variables
by State, Year and Month.

State 4 — Arizona

	Month		
Year	8	9	10
80	BOLLUN TOTBL	TOTBL WTBOLL	CUMWT BOLLUN
81	BOLLUN BOLLOP	TOTBL WTBOLL	TOTBL WTBOLL
82	TOTBL BOLLOP	TOTBL WTBOLL	TOTBL CUMWT
83		no data	
84	TOTBL BOLLOP	TOTBL WTBOLL	TOTBL WTBOLL
85	TOTBL BOLLOP	TOTBL WTBOLL	TOTBL CUMWT
86	TOTBL BOLLOP	TOTBL WTBOLL	TOTBL WTBOLL

State 5 — Arkansas

	Month		
Year	8	9	10
80	TOTBL TOTSQ	TOTBL CUMWT	BOLLOP BOLLUN
81	TOTBL TOTSQ	TOTBL BOLLOP	CUMWT TOTBL
82	TOTBL TOTSQ	TOTBL CUMWT	CUMWT TOTBL
83	TOTBL TOTSQ	TOTBL BOLLOP	CUMWT TOTBL
84	TOTBL TOTSQ	TOTBL BOLLOP	CUMWT TOTBL
85	TOTSQ TOTBL	TOTBL BOLLOP	CUMWT BOLLUN
86	TOTBL TOTSQ	TOTBL BOLLOP	CUMWT TOTBL

Table C2 (Continued)

State 6 — California

	Month					
Year	8		9		10	
80	BOLLUN	TOTBL	TOTBL	BOLLOP	TOTBL	WTBOLL
81	BOLLUN	TOTBL	TOTBL	PLT	CUMWT	BOLLUN
82	BOLLUN	TOTBL	TOTBL	PLT	CUMWT	BOLLUN
83	BOLLUN	TOTBL	TOTBL	PLT	TOTBL	CUMWT
84	BOLLUN	TOTBL	TOTBL	PLT	TOTBL	CUMWT
85	BOLLUN	TOTBL	TOTBL	PLT	TOTBL	CUMWT
86	BOLLUN	TOTBL	TOTBL	PLT	CUMWT	BOLLUN

State 22 — Louisiana

	Month					
Year	8		9		10	
80	TOTSQ	MATUR	TOTBL	CURWT	CUMWT	BOLLUN
81	TOTSQ	MATUR	TOTBL	CURWT	CUMWT	BOLLUN
82	TOTSQ	BOLLUN	TOTBL	CURWT	CUMWT	BOLLUN
83	TOTSQ	MATUR	TOTBL	CURWT	CUMWT	BOLLUN
84	BOLLUN	TOTSQ	TOTBL	TOTSQ	CUMWT	BOLLUN
85	TOTSQ	BOLLUN	TOTBL	CURWT	CUMWT	BOLLUN
86	TOTSQ	MATUR	TOTBL	CURWT	CUMWT	BOLLUN

Table C2 (Continued)

State 28 — Mississippi

	Month					
Year	8		9		10	
80	TOTSQ	TOTBL	TOTBL	CURWT	CUMWT	BOLLUN
81	TOTSQ	TOTBL	TOTBL	CURWT	CUMWT	BOLLUN
82	TOTSQ	TOTBL	TOTBL	CURWT	CUMWT	BOLLUN
83	TOTSQ	TOTBL	TOTBL	CURWT	CUMWT	BOLLUN
84	TOTBL	TOTSQ	TOTBL	BOLLUN	CUMWT	BOLLUN
85	TOTSQ	BOLLUN	TOTBL	CURWT	CUMWT	BOLLUN
86	TOTSQ	TOTBL	TOTBL	BOLLOP	CUMWT	BOLLUN

State 48 — Texas Region 1

	Month					
Year	8		9		10	
80	CUMWT	CURWT	BOLLOP	WTBOLL	BOLLOP	WTBOLL
81	TOTBL	BOLLOP	BOLLOP	WTBOLL	BOLLOP	WTBOLL
82	CUMWT	CURWT	BOLLOP	WTBOLL	BOLLOP	WTBOLL
83	CUMWT	CURWT	BOLLOP	WTBOLL	BOLLOP	WTBOLL
84	CUMWT	CURWT	BOLLOP	WTBOLL	BOLLOP	WTBOLL
85	CUMWT	CURWT	CUMWT	BOLLUN	BOLLOP	WTBOLL
86	TOTBL	BOLLOP	BOLLOP	WTBOLL	BOLLOP	WTBOLL

Table C2 (Continued)

State 48 — Texas Region 2

	Month					
Year	8		9		10	
80	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	BOLLUN
81	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	BOLLUN
82	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	BOLLUN
83	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	BOLLUN
84	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	WTBOLL
85	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	BOLLUN
86	TOTSQ	BOLLUN	TOTBL	TOTSQ	TOTBL	PLT

LEGEND: Most of the variables in Table C2 are defined in Table C1. However, the following two variables are not:

1.) CURWT = weight of harvested bolls during current month.

2.) MATUR = "maturity index" which is defined as number of bolls/plant, *i.e.* TOTBL/PLT.

Appendix D – Current NASS Forecast Procedures

This appendix summarizes the current NASS cotton objective yield forecast and estimation procedure. It contains a slightly edited version of only those parts of the NASS Objective Yield Manual that are essential to specify the current cotton yield forecast and estimation procedures. For further information the reader should consult the NASS Objective Yield Manual, May 1989, Section 8.5, Forecasting and Estimating Models, Cotton Forecasting and Estimation, pages 8501–8520, .

I. General

Sample fields for the Cotton Objective Yield Survey are selected from farms reporting upland cotton for harvest in the AREA frame of the June Agricultural Survey (JAS). The Objective Yield Survey is conducted in six states. In California and Texas, the sample is selected and summarized in two distinct Districts. The other four states are each treated as a District. Each month during the Objective Yield Survey, data collected from the sample fields are used to produce indications (forecasts or estimates) of at harvest yield as specified below.

II. Yield

A. Maturity Categories

Yield is forecast for each sample according to its maturity category, which is determined within the summary program as follows.

Maturity	Fruit Counted	Fruit Present
Category	Within Units	Beyond Units
1	No fruit present	No fruit present
2	No fruit present	Squares only
3	$0 \leq$ RATIO < 0.5	Blooms or Bolls
4	$0.5 \leq$ RATIO < 2.0	—
5	$2.0 \leq$ RATIO	—
6	Sample field has been harvested since the initial Form-B was completed.	

Note: RATIO is the ratio of large bolls counted to plants counted in the 10-foot units. Large bolls include burrs, open bolls, partially open bolls, and large unopened bolls.

B. Estimating and Forecasting Procedures

The objective yield samples are selected in such a way that each acre has equal probability of selection within Districts. Therefore, the average of the sample level yields across all samples in a District provides a forecast of mean gross yield per

acre for the District. Also, the average harvest loss, as determined from Form-E data, provides a District level estimate of harvest loss. If fewer than 10 Form-E's (post-harvest field observation data recording forms used in the estimation harvest loss) have been completed within a District, a five-year average harvest loss is used.

Gross Yield

The final estimate of gross yield is computed by multiplying the number of large bolls at harvest by the average weight per boll, expanding to a "per acre" basis, and converting to a standard unit. The standard unit for cotton is "pounds of lint at 5 percent moisture". Production is reported in 480-pound bales.

The formula for computing gross yield is:

$$GY = (2.401 \times LSR \times LB \times BW)/RS$$

where,

GY = Gross Yield (in lbs. of lint per acre)
LSR = Lint/Seed Ratio (3-year average)
LB = Number of large bolls at harvest (on a 40-foot basis)
BW = Average boll weight (in grams at 5 percent moisture, gin equivalent)
2.401 = $43,560/(40 \times 453.59)$ which converts grams of seed cotton per 40 feet of row to pounds of seed cotton per acre.

Harvest Loss

The harvest loss is computed from gleanings obtained in all usable even-numbered samples. The sample level harvest loss is found by determining the total weight of seed cotton gleaned, expanding to a "per acre" basis, and converting to standard units.

The formula for harvest loss is:

$$HL = (2.401 \times WT \times LSR)/RS$$

where,

HL = Harvest Loss (lbs. of lint per acre)
WT = Weight of cotton left in units, which is computed as the, (partially opened and large unopened bolls left in the units) × (average net weight per boll) + (weight of cotton gleaned adjusted to 5 percent moisture).
LSR = Lint/Seed ratio
RS = Row space measurement
2.401 = Conversion factor as defined above.

Early in the growing season, some or all of the three components of net yield (number of bolls, average boll weight, and harvest loss) cannot be obtained directly and must be forecast. The procedures used to forecast these components are described in the following sections.

C. **Forecasting the Number of Large Bolls**

The expected number of large bolls for each sample is forecast using a survival model and a regression model:

$$\text{SURVIVAL: } Y = YTC + (LB \times X1) + (SBBL \times X2) + (SQ \times X3)$$
$$\text{REGRESSION: } Y = B1 + (B2 \times X1) + (B3 \times X2) + (B4 \times X3)$$

where,

Y	=	Forecasted number of large bolls
$X1$	=	Observed number of burrs, open bolls, partially open bolls, and large unopened bolls (40-foot equivalent)
$X2$	=	Observed number of small bolls and blooms (40-foot equivalent)
$X3$	=	Observed number of squares (40-foot equivalent)
YTC	=	Large bolls "yet to come" from fruit not yet set
LB	=	Survival ratio of all large bolls
$SBBL$	=	Survival ratio of small bolls and blooms
SQ	=	Survival ratio of squares
$B1$–$B4$	=	Least squares regression coefficients.

Forecast equations for each model are derived for each maturity category for each month for each District for each State, resulting in 336 possible forecast equations. Data from the previous five years are used to derive the survival ratios and regression coefficients. If a unique set of coefficients cannot be determined for a given class (due to lack of sufficient data), the previous month's coefficients are used.

The actual count of large bolls is used for any sample in maturity category six in any month, and for all samples in December and later months. All samples in maturity category one use a five-year historic average.

The survival equations are derived from data collected in the 3-foot tag sections adjacent to Unit 1 of each sample. At the end of each survey year, tag data is summarized to calculate survival ratios by classes. These ratios are added to a "record book" containing survival ratios from previous years. This record book is used to compute the five-year averages. (Some classes may not have data for some years, so the five-year averages for these classes will actually go back more than calendar five years.) The constant YTC (bolls "yet to come") is computed only for August classes since it is assumed that fruit will have set by the September survey period.

The regression equations are derived from the previous five years' survey data using multiple regression techniques. Certain influential data points (i.e., "outliers") are excluded from the dataset prior to deriving the coefficients. These influential data points are identified using a "deleted residual" analysis or the "Cook's D" statistic. There is usually very little change in the regression equations from year-to-year because roughly 80 percent of the data for each class was used in the analysis the previous year. Classes that do change significantly from one year to the next are usually those with very few observations. If a class has little data and a plausible forecast equation cannot be derived, the equation from the previous year is used.

The two forecasts (survival and regression) are weighted by the inverse of their forecast errors to form a composite forecast. Forecast error for each model is calculated from the five-year dataset used to compute the regression coefficients and is defined as:

$$FE(i) = \left[\frac{\sum(Y(i,j) - \hat{Y}(i,j))^2}{n}\right]^{1/2}$$

where,

$\hat{Y}(i,j)$ = Forecast number of large bolls for model i and sample j
$Y(i,j)$ = Final number of large bolls for model i and sample j
n = Number of samples for a maturity category within a District
$FE(i)$ = Forecast error for maturity category i (either survival or regression)

The inverse of the forecast error is a weighting factor which gives larger weight to the model with the smaller forecast error.

D. Forecasting Boll Weight

One model is used to forecast boll weight for all maturity categories in a District in a State. Early in the year (until 20 percent of the projected number of large bolls have been picked and weighed by the enumerator) a five-year historic average is used. The following model is used during the season, when between 20 and 85 percent of the projected number has been picked and weighed:

$$Y = W \times (A + BX)$$

where,

A & B are fixed regression coefficients from the table below
Y = Forecast boll weight
W = Observed boll weight to date
X = Ratio of bolls picked and weighed to large bolls forecasted.

When more than 85 percent of the projected number of large bolls has been picked and weighed by the enumerator, and until the field is harvested, actual boll weight is used.

State	*A*	*B*
Arizona	.863	.153
Arkansas	.863	.153
California		
San Joaquin	.945	.058
Imperial Valley	.882	.131
Louisiana	.882	.131
Mississippi	.882	.131
Texas		
East	.921	.086
West	.883	.137

III. Example: Yield Computed for a Single Sample

A. Assume the following September 1 Data

8-row space measurement (no skip) 25.8

Counts Within 10-foot Units

Number of plants (4 rows) 87
Number of burrs (2 units) 113
Total open bolls (4 rows) 130
Weight of seed cotton picked (2 units) 650
Number of partially open bolls (4 rows) 48
Number of large unopened bolls (4 rows) 121

3-foot Tag Section Beyond Unit 1

Number of plants 11
Number of burrs and open bolls
- red tags 6
- yellow tags 21
- no tags 6

Number of large unopened bolls
- red tags 0
- yellow tags 12
- no tags 2

Number of small bolls and blooms

yellow tags 1
blue tags 3
Number of squares 2

3-foot Count Section Beyond Unit 2

Number of plants 8
Number of burrs and open bolls 27
Number of large unopened bolls 11
Number of small bolls and blooms 6
Number of squares 1

Current Month Lab Form

Weight of seed cotton before drying 56
Weight of seed cotton after drying 52

Previous Months' Data Brought Forward

Accumulated burrs within unit 20
Accumulated bolls picked within unit 50
Accumulated adjusted weight seed cotton 257

B. Maturity Category Determination

$$\begin{aligned} LB &= \text{Burrs + open bolls + partially open bolls} \\ &\quad + \text{large unopened bolls within unit} \\ p &= \text{Number of plants} \\ LB/p &= (113 + 20) + (130 + 50) + 48 + 121/87 \\ &= 5.54 \end{aligned}$$

So the maturity category is 5 (ratio > 2.00).

C. Forecast Number of Bolls

1. Multiple Regression Model

$$NB(r) = \#\text{ bolls } = B1 + (B2 \times X1) + (B3 \times X2) + (B4 \times X3)$$

where,

$B1 = 14$ regression
$B2 = .933$ coefficients derived
$B3 = .300$ from previous five
$B4 = .110$ years of data
$X1$ = Burrs and large bolls (40-ft equivalent)
$X2$ = Small bolls and blooms (40-ft equivalent)
$X3$ = Square (40-ft equivalent.)

Since burrs and open bolls and partially open bolls and large unopend bolls are counted in a total of 46 feet of row (four 10-foot units and two 3-foot units),

$$\begin{aligned} X1 &= (40/46) \times (\text{all large bolls}) \\ &= (40/46) \times ((113+20)+(130+50)+48+121+(33+14)+(27+11)) \\ &= (40/46) \times 567 \\ &= 493.043 \end{aligned}$$

Since small bolls and blooms are counted in six feet of row (both 3-foot units),

$$\begin{aligned} X2 &= (40/6) \times (4+6) \\ &= 66.67 \end{aligned}$$

Since squares are counted in six feet of row (both 3-foot units),

$$\begin{aligned} X3 &= (40/6) \times (2+1) \\ &= 20.001 \end{aligned}$$

So, the estimate of number of bolls using the regression model for this sample is:

$$\begin{aligned} NB(r) &= 14 + (.933 \times 493.043) + (.300 \times 66.67) + (.110 \times 20.001) \\ &= 496.210 \end{aligned}$$

2. Survival Model

$$NB(s) = \#\text{ bolls } = YTC + (LB \times X1) + (SBBL \times X2) + (SQ \times X3)$$

where the X's are defined as in the regression models and,

$$\begin{aligned} YTC &= \text{Number of large bolls yet to come} \\ &\quad \text{from fruit set after September 1} \\ LB &= \text{Large boll survival rate} \\ SBBL &= \text{Small boll and bloom survival rate} \\ SQ &= \text{Square survival rate.} \end{aligned}$$

These coefficients are derived from analysis of tag section data. The expected number of large bolls yet to come from fruit set after August 1 and the expected monthly survival of large bolls, blooms and small bolls, and squares are computed from each of the last five years. The average survival rates are used as coefficients. For this example, let

$$\begin{aligned} YTC &= 0 \\ LB &= .919 \\ SBBL &= .315 \\ SQ &= .105 \end{aligned}$$

So, the estimate of number of bolls using the survival model is:

$$\begin{aligned} NB(s) &= 0 + (.919 \times 493.043) + (.315 \times 66.67) + (.105 \times 20.001) \\ &= 476.208 \end{aligned}$$

3. Combining the Two Forecasts

The two forecasts of number of bolls (one from the regression model and the other from the survival model) are weighted together to obtain a combined forecast. The weight, which were defined in Section II.C are inverse of the forecast error. Thus the weights are given by:

$$\begin{aligned} W(r) &= \text{Weight of regression model forecast} \\ &= FE(s)/(FE(s)+FE(r)) \\ &= \\ W(s) &= \text{Weight of survival model forecast} \\ &= FE(r)/(FE(s)+FE(r)) \end{aligned}$$

where,

$$\begin{aligned} FE(s) &= \text{Survival model forecast error} \\ FE(r) &= \text{Regression model forecast error.} \end{aligned}$$

In this example, let

$$\begin{aligned} W(r) &= .521 \\ W(s) &= .479. \end{aligned}$$

Then, the combined forecast of number of bolls is:

$$\begin{aligned} NB &= W(r) \times NB(r) + W(s) \times NB(s) \\ &= .521 \times 496.210 + .479 \times 476.208 \\ &= 486.629 \end{aligned}$$

D. Forecast Boll Weight

$$BW = Z21 \times (A + B \times Z22)$$

where,

A & B are regression coefficients

$Z21$ = Accumulated weight of seed cotton picked (adjusted for moisture content) divided by the accumulated number of open bolls picked

$Z22$ = Accumulated number of open bolls picked divided by the forecast number of large bolls (i.e., this is the proportion of forecast large bolls picked by the enumerator.)

For this example, let:

$$\begin{aligned} A &= .882 \\ B &= .131 \end{aligned}$$

To determine $Z21$ for the current month:

$$\begin{aligned}\text{Drying ratio} &= \text{Dry weight / Wet Weight}\\ &= 52/56\\ &= .9286\end{aligned}$$

$$\begin{aligned}\text{Current month's weight picked} &= 650 \times .9286\\ &= 603.590\end{aligned}$$

$$\begin{aligned}\text{Current month's weight (at 5\% moisture)} &= 603.590 \times 1.0526\\ &= 635 \text{ grams.}\end{aligned}$$

Since 1.0526 is the conversion factor to 5% moisture (gin equivalent):

$$\begin{aligned}Z21 &= (257 + 635)/(50 + 130)\\ &= 4.956\end{aligned}$$

$$\begin{aligned}Z22 &= 180/487\\ &= .370\end{aligned}$$

and

$$\begin{aligned}BW &= Z21 \times (A + B \times Z22)\\ &= 4.956 \times (.882 + .131 \times .370)\\ &= 4.611 \text{ grams per boll.}\end{aligned}$$

E. Forecast Gross Yield per Acre

Using the formula in Section II.B, the estimated gross yield for this sample is:

$$\begin{aligned}GY &= (2.401 \times LSR \times NB \times BW)/RS\\ &= (2.401 \times .368 \times 486.629 \times 4.611)/3.225\\ &= 614.76 \text{ pounds of lint per acre.}\end{aligned}$$

The average estimated gross yield across all samples in a District less an estimate of harvest loss produces the District estimate of Net Yield.

Appendix E – The Objective Yield Sample by Maturity Category.

TABLE E1

The Percentage of 1982-1986 Cotton Objective Yield Sample Plots Falling in the Various Maturity Categories in August, September and October along with the Mean Monthly NASS and Markov Forecast Error.

State 4 — Arizona

Category	August	September	October
1	0.40	.	.
2	3.43	.	.
3	23.19	1.02	.
4	20.77	3.07	.
5	52.22	95.91	99.80
6	.	.	0.20
Mean NASS Forecast Error	5.6	2.8	3.8
Mean Markov Forecast Error	2.8	2.4	1.9

State 5 — Arkansas

Category	August	September	October
1	1.03	.	.
2	5.95	.	.
3	51.95	2.24	0.20
4	26.69	13.27	7.96
5	14.37	84.49	90.82
6	.	.	1.02
Mean NASS Forecast Error	13.8	13.3	13.8
Mean Markov Forecast Error	10.9	10.1	11.0

TABLE E1 (Continued)

State 6 — California

Category	August	September	October
1	0.39	.	.
2	2.58	0.16	.
3	27.39	0.62	0.31
4	30.91	2.33	0.70
5	38.65	96.81	98.91
6	0.08	0.08	0.08
Mean NASS Forecast Error	3.8	5.1	1.5
Mean Markov Forecast Error	5.5	5.1	4.7

State 22 — Louisiana

Category	August	September	October
1	0.85	0.22	.
2	11.11	.	.
3	35.04	1.10	.
4	33.97	8.81	1.93
5	19.02	89.87	94.42
6	.	.	3.65
Mean NASS Forecast Error	17.7	13.7	4.4
Mean Markov Forecast Error	18.5	13.1	17.8

TABLE E1 (Continued)

State 28 — Mississippi

Category	August	September	October
1	0.39	.	.
2	5.10	.	.
3	41.62	0.65	0.13
4	29.71	5.76	1.18
5	23.17	93.59	94.11
6	.	.	4.58
Mean NASS Forecast Error	17.8	8.3	4.9
Mean Markov Forecast Error	15.8	16.1	17.3

State 48 — Texas, Region 1

Category	August	September	October
1	0.27	.	.
2	2.18	.	.
3	10.08	2.12	0.27
4	16.35	15.65	7.22
5	71.12	60.48	28.34
6	.	21.75	64.17
Region 1 and 2 Combined			
Mean NASS Forecast Error	12.7	8.8	4.2
Mean Markov Forecast Error	2.9	2.0	2.8

TABLE E1 (Continued)

State 48 — Texas, Region 2

Category	August	September	October
1	15.53	0.74	0.05
2	50.27	2.52	.
3	33.10	30.50	3.33
4	0.79	44.67	38.16
5	0.30	21.57	58.37
6	.	.	0.10
Region 1 and 2 Combined			
Mean NASS Forecast Error	12.7	8.8	4.2
Mean Markov Forecast Error	2.9	2.0	2.8

All States

Category	August	September	October
1	5.67	0.27	0.02
2	20.27	0.90	.
3	32.41	11.21	1.25
4	18.55	19.68	14.52
5	23.08	66.53	79.08
6	0.02	1.41	5.14
Mean NASS Forecast Error	11.3	8.2	4.5
Mean Markov Forecast Error	7.1	6.1	6.9
Mean Percent of Bolls Picked	2.2	7.9	41.4

The mean forecast errors in Table E1 are weighted means with weights proportional to the October sample size. The mean percent of bolls picked by all enumerator are relative to the total number of boll forecast by the NASS regression model over all states for the years 1982–1986.

Table E2

Mean over all States of Cumulative Percentage of Forecast Bolls Picked by the Enumerator through the Specified Month.

YEAR	August	September	October
1982	1.3	5.7	36.4
1983	1.1	4.4	41.8
1984	3.5	6.7	36.7
1985	2.2	7.5	46.6
1986	3.1	15.3	46.3
Mean 1982–86	2.2	7.9	41.4

Table E3

The Percentage of the 1982–86 Objective Yield Sample Plots Falling in the Various States and Regions within States.

state	August	September	October
Arizona	8.44	8.31	8.37
Arkansas	8.29	8.33	8.37
California Region 1	19.37	19.44	19.59
California Region 1	2.38	2.40	2.39
California Region 1 and 2	21.75	21.84	21.98
Louisiana	7.97	7.71	7.96
Mississippi	13.00	12.98	13.05
Texas Region 1	6.25	6.41	6.39
Texas Region 2	34.30	34.43	33.89
Texas Region 1 and 2	40.54	40.83	40.27

Appendix F – Modification to Ensure Convergence of Markov Forecast

Even though the Markov procedure appears to be superior to the current NASS forecast procedure early in the season, it has two undesirable characteristics. Unlike the current NASS forecast procedure, the Markov forecast neither tends to improve as the season progresses nor does it tends to converge at the end of the growing season to the final objective yield estimate.

Combining the NASS forecast equations for boll number and average boll weight and evaluating the parameters in the resulting equation at their idealized at–harvest values produces essentially the final objective yield estimation equation. In other words, as the growing season progresses the combined NASS forecast procedure gets closer to the final objective yield estimate. In contrast, as was observed in the body of this report, the Markov forecast model always relies on historical yield values for its forecast so the final Markov forecast model does not necessarily approach the final objective yield estimation model. In other words, the final Markov forecast is always the average of historical plot yield values and thus never approaches the average of current plot yield values.

To overcome the difficulties referred to above, modifications to the current Markov procedure are recommended. The basic idea is to use the Markov procedure to forecast the gross yield remaining in the plot after the current data collection period. And then combine this forecast with the cumulative enumerator harvested yield to obtain an at–harvest gross yield forecast. This modified Markov procedure will converge to the final objective yield estimate as one moves through the growing season and as percent of the plot harvested by the enumerator increases.

Let the residual yield, the Markov residual forecast yield, the cumulative enumerator harvested yield, the final at–harvest objective yield and the forecast at–harvest objective yield be denoted by Y_R, y_R, Y_H, Y_F and y_F respectively. In terms of the yield components, the model for the at–harvest forecast can be written as

$$y_F = Y_H + y_R + e_F$$

where e_F is the error of the forecast and $Y_F = Y_H + Y_R$.

It is probably reasonable to assume that the investigated and modified Markov procedures have about the same relative forecast errors since both procedures are the same except that one procedure is based on historical total season plot yields and the other procedure is based on historical yield collected from the plot after the forecast month. This assumption is further supported by the fact that the relative error of the investigated Markov procedure does not appear to improve as the growing season progresses.

With these assumptions a rough estimate of the forecast error associated with the modified forecast procedure is derived as follows.

$$\begin{aligned}\frac{E\{[Y_F - y_F]^2\}}{F^2} &= \frac{E\{[Y_F - (Y_H + y_R)]^2\}}{F^2} \\ &= \frac{E\{[(Y_F - Y_H) - y_R)]^2\}}{F^2} \\ &= \frac{E\{[Y_R - y_R]^2\}}{F^2} \\ &= \frac{R^2}{F^2} \times \frac{E\{[Y_R - y_R]^2\}}{R^2}\end{aligned}$$

For a specified forecast month the ratio R/F can be approximated by one minus the average percent of bolls picked by the enumerator. Assuming the values given in Appendix E, which are repeated in part below, a rough idea of the forecast error associated with the modified Markov process is obtained as listed in the last line of the table below.

MEAN NASS FORECAST ERROR	11.3	8.2	4.5
MEAN MARKOV FORECAST ERROR	7.1	6.1	6.9
MEAN PERCENT OF BOLLS PICKED	2.2	7.9	41.4
MEAN PERCENT OF BOLLS LEFT	97.8	92.1	58.6
MODIFIED MARKOV FORECAST ERROR	7.0	5.6	4.0

Under the assumptions given above it would appear that the forecast error of a modified Markov procedure would be smaller then the forecast error of either the NASS procedure or the investigated Markov procedure.

Appendix G – Time and Cost Estimates for the Proposed Research

The thirteen specific recommendations for future research were grouped into four general research areas, or parts, in the body of the report. This appendix contains a brief summary of each part together with an associated time and cost estimate. Each time frame and budget indicates the actual time and cost estimated to accomplish the proposed research work. Thus, the work proposed for one year could be completed over multiple calendar years as funding is available. Part 1 has the highest priority and is recommended for immediate funding. In addition, it would be very cost efficient to commence one or more of the other parts, with at least some concurrent funding, to capitalize on the synergism involved with multiple parts.

Part 1. Modification of Present Markov Chain Procedure and Comparison to Current NASS Forecast Procedure.

Objectives:

1. Transform dependent variable to ensure convergence to final objective yield estimate.
2. Investigate possible contribution of other independent variables (*e.g.* pasture and range condition data).
3. Investigate the effect of changing the number of intermediate states.
4. Prepare production mode P.C. software for Markov chain forecast.
5. Implement NASS forecast procedures and use bootstrap techniques to compare NASS and Markov chain forecasts.
6. Assess and compare forecast errors of each of the two component parts of NASS procedure, *i.e.* regression model and survival model forecasts.
7. Develop a similarity index.

Estimated Time Required for Part 1. It is expected that most, if not all, of the research outlined in Part 1 could be completed within one research–year with the following proposed budget.

Proposed Budget for Part 1.

Salaries:	
Professional 4 mo. @ 6,000 =	$24,000
Graduate Assistant 1 yr. @ 12,000 =	12,000
Clerical 2 mo. @ 1,500 =	3,000
Computer Programming 2 mo. @ 1,500 =	3,000
Travel	1,800
Publication Costs	700
Computer Services	4,000
Total	$48,500

Part 2. Develop Markov Process Models to Describe and Utilize Tag Survival Data.

Objectives:

1. Formulate and fit semi–Markov process models to plot level survival (tag) data.
2. Assess the value of tag data and, if cost effective, incorporate the tag data into new optimal forecasting methods.

Estimated Time Required for Part 2. This part is estimated to be a two research–year project. The project could be initiated with the following start–up costs for an exploratory analysis of the tag data. A concentrated modeling effort would follow in the second research–year.

Proposed Additional Budget for Part 2:

Salaries:	
Professional 2 mo. @ 6,000 =	$12,000
Graduate Assistant 1/2 yr. @ 12,000 =	6,000
Travel	600
Publication Costs	400
Computer Costs	2,500
Total	$21,500

Part 3. Extensions of Software and Generalizations of Markov Chain Procedure.

Objectives:

1. Develop completely integrated, one–step PC software.
2. Investigate use of multivariate methods to define states.

Estimated Time Required for Part 3. This part of the research should be completed within a one research–year period by a graduate assistant.

Proposed Additional Budget for Part 3:

Salaries:	
Graduate Assistant 1 yr. @ 12,000 =	$12,000
Travel	600
Publication Costs	400
Computer Costs	1,000
Total	$14,000

Part 4. Investigate Use of Markov Forecast Procedures for Other Crops.

Estimated Time Required for Part 4. It is obviously difficult to estimate how long this aspect of the research would last. No doubt, the duration of the research would be a function of its practical utility. The following budget gives the estimated additional first research–year start–up costs for this part.

Salaries:	
Professional 1 mo. @ 6,000 =	$6,000
Graduate Assistant 1/2 yr. @ 12,000 =	6,000
Travel	600
Publication Costs	400
Computer Costs	1,000
Total	$14,000

* U. S. GOVERNMENT PRINTING OFFICE:1989-261-493:22325/case